The International Conference on Management, Business, Economics, Law and Technology (COMBELT-2025)

August 01st, 2025, Danang, Vietnam

From Black-Box to Glass-Box: Explainable Deep Learning for Credit Risk Prediction

Chau Thi Da Huong[a]✉, Le Dien Tuan[a]

[a] *The University of Danang – University of Economics, 71 Ngu Hanh Son, Danang 550000, Vietnam*

ABSTRACT

Accurate and interpretable credit scoring is essential for institutions of finance to assess applicants' creditworthiness and support informed lending decisions. While deep learning models have proven strong predictive capacities in credit risk assessment, their black-box nature often hinders interpretability and practical deployment. This study investigates the effectiveness of several models, including Random Forest, XGBoost, LightGBM, and Artificial Neural Network (ANN) on the German Credit dataset to predict whether a loan applicant poses a good or bad credit risk. According to experimental findings, the ANN model consistently outperforms others across evaluation metrics, achieving the highest accuracy of 0.7250, F1 score of 0.6742, and AUC of 0.7606 on the test set. To address the interpretability challenge, we incorporate explainable AI (XAI) methods, namely SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), to provide both global and local interpretations of the model's outputs. The consistency between SHAP and LIME enhances confidence in model decisions, offering actionable insights for stakeholders. This research emphasizes the possibility of combining high-performing deep learning models with interpretable AI for responsible, transparent and effective credit risk assessment.

Keywords: credit scoring; credit risk prediction; deep learning; machine learning; artificial neural network; explainable AI.

1. Introduction

Durand (Durand 1941) first introduced credit scoring, which has evolved into a vital risk management method for lending organizations. It provides a quantitative assessment of an individual's creditworthiness, enabling financial institutions to evaluate repayment ability and

✉ Corresponding author. Tel.: +84 975 751 421.
Email address: huongctd@due.edu.vn

make informed lending decisions. This not only benefits individuals through improved access to credit but also empowers businesses to manage financial risks and credit allocation more effectively.

Traditional credit scoring systems, often based on logistic regression or rule-based approaches, offer interpretability but struggle to capture complex, nonlinear patterns in borrower behavior. The development of advanced ensemble models, such as feature-optimized and deep learning-based networks, has led to innovation in financial risk estimation since they provide deeper insights into risk and perform well with large and high-dimensional datasets. Deep learning outperforms conventional machine learning techniques in several areas of credit risk assessment (Shen et al. 2021).

Nevertheless, the enhanced accuracy of these models often comes at the expense of transparency. One major obstacle to adoption in regulated environments is the "black-box" aspect of advanced machine learning, especially deep learning models. Stakeholders, including credit officers, regulators, and customers, need to understand not only what decisions are made but also why they are made.

This research aims to create an interpretable deep learning framework, addressing the dual challenge of performance and interpretability in credit scoring. Specifically, we assess and compare the predictive effectiveness of four widely-used models namely Random Forest, XGBoost, LightGBM, and Artificial Neural Network (ANN) in credit risk prediction. To tackle the interpretability issue, we incorporate two XAI techniques including SHAP and LIME. These methods provide global and instance-level explanations regarding the ANN model's predictions, enhancing transparency and trust in model decisions.

The rest of this study is structured as follows: Section 2 reviews various related works. Section 3 outlines the proposed approach used in this study. Section 4 discusses and analyzes the experiment findings. Finally, Section 5 wraps up the work and offers suggestions for further research.

2. Related Work

The application of machine learning and deep learning approaches to credit scoring has gained more attention due to its capacity to automatically extract pertinent information from large data sets and identify complex patterns. Chen et al. (Chen & Zhang 2021) addressed the challenge of imbalanced credit card default data by proposing a predictive model that combines k-means SMOTE with a Backpropagation Neural Network (BPnn). In their approach, the k-means SMOTE algorithm is used to rebalance the data distribution, while a Random Forest model identifies important features, which are then incorporated into the initial weight setting of the BPnn. The performance of this method was compared against five conventional machine learning models including logistic regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and decision trees. Museba (Museba 2022) proposed a heterogeneous adaptive particle swarm optimization method, further showcasing the effectiveness of metaheuristic-based techniques on credit scoring. Ensemble learning approaches, particularly those based on XGBoost, have consistently demonstrated strong performance in credit risk prediction.

Deep learning models have made significant progress in the field because of their non-linear learning abilities and exceptional feature extraction. For instance, Luo et al. (Luo et al. 2017) illustrated the effectiveness of Deep Belief Networks (DBNs), which outperformed traditional

classifiers like multilayer perceptrons (MLPs), logistic regression and SVMs. Zhu et al. (Zhu et al. 2018) proposed a hybrid approach that integrates the Relief feature selection algorithm with a Convolutional Neural Network (CNN), resulting in a Relief-CNN model that surpassed traditional models like Logistic Regression and Random Forest in predictive accuracy. Additionally, Gicić et al. (Gicić et al. 2023) employed stacked unidirectional and bidirectional LSTM (Long Short-Term Memory) networks to tackle credit scoring tasks, achieving notable performance improvements across multiple benchmark credit datasets.

These studies collectively highlight the growing importance and effectiveness of advanced machine learning and deep learning models in improving the accuracy and robustness of credit risk prediction. However, the opacity of these models has raised significant concerns, particularly in high-stakes domains such as finance. To bridge the interpretability gap, Explainable AI (XAI) has emerged as a promising solution that enables stakeholders to understand, trust, and validate complex model predictions.

Several XAI methods, including SHapley Additive exPlanations (SHAP) (Lundberg & Lee 2017) and local interpretable model-agnostic explanations (LIME) (Ribeiro et al. 2016), are frequently employed to enhance the transparency. Talaat et al. (Talaat et al. 2024) proposed CreditNetXAI, a hybrid framework that integrates deep learning with model-agnostic explanation techniques, specifically SHAP, to enhance both predictive performance and interpretability in credit card default prediction. Similarly, Zou et al. (Zou et al. 2025) introduced Add-XGBoost, an ensemble model that combines the interpretability of Generalized Additive Models (GAM) with the power of Extreme Gradient Boosting. Add-XGBoost demonstrated superior performance, offering both global and local interpretability by revealing the interactions and individual contributions of features to the final credit risk predictions. These studies underscore the growing importance of combining high-performing predictive models with transparent, interpretable decision-making frameworks in the credit scoring domain.

3. The Proposed Approach

The research process is divided into four key phases including data collection, data preprocessing, model training and validation, evaluation and interpretation, as shown in Fig. 1.

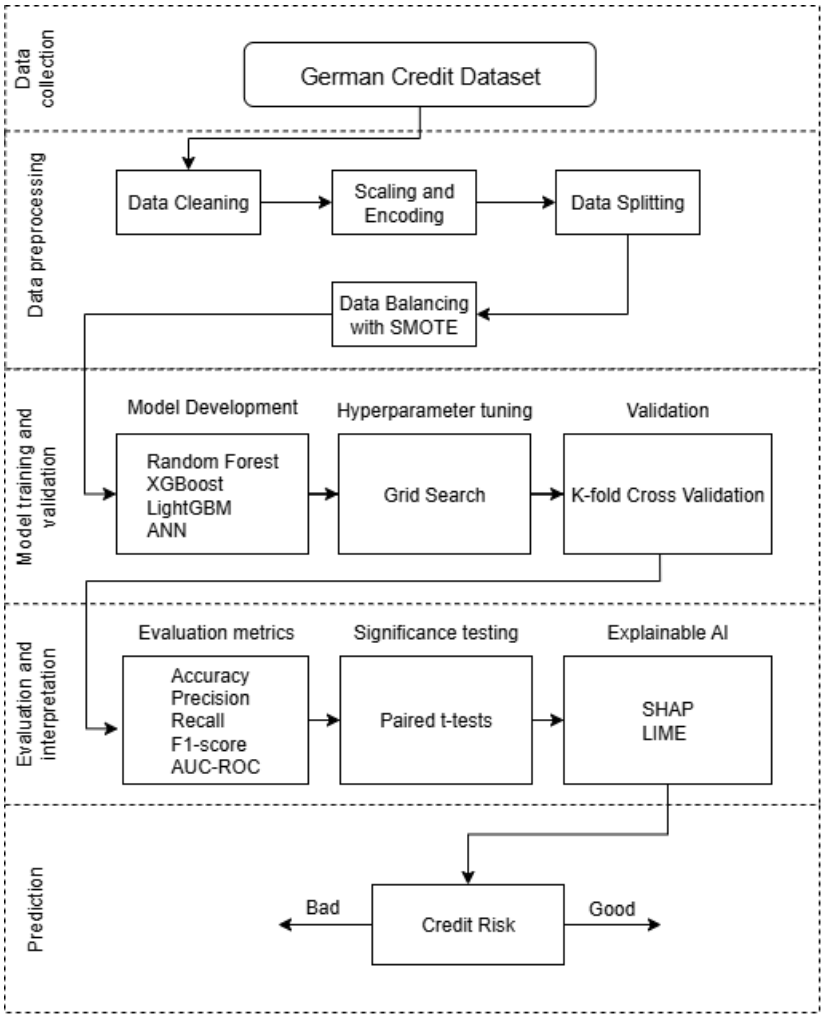

Fig. 1. Proposed research framework.

3.1. Data Collection

The research utilizes the publicly available German credit dataset, originated from UCI Machine Learning Repository. It includes 1000 records of credit from a German bank, each labeled as having either good or bad credit risk. This dataset contains 20 features encompassing both numerical and categorical variables, such as loan duration, credit amount, employment status, personal information, and financial history, making it a widely used benchmark for evaluating credit scoring models.

3.2. Data Preprocessing

To make sure the dataset was balanced, clean, and ready for model training, several preprocessing steps were applied. First, missing values, duplicate instances, and outliers were checked, and the data was standardized to ensure consistent formatting. Next, numeric features were normalized using Z-score standardization, while categorical variables were transformed using one-hot encoding.

Following preprocessing, the dataset is split into two sections: 20% for testing and 80% for training to evaluate model generalization. Since the dataset exhibited a class imbalance, containing 300 instances of bad credit and 700 instances of good credit, the Synthetic Minority Oversampling

Technique (SMOTE) was used to create synthetic samples for the minority class on the training set, improving class balance and helping prevent model bias toward the majority class.

3.3. Model Training and Validation

Multiple classification models were developed, including XGBoost, Random Forest, LightGBM, and a deep ANN, to facilitate a comparative performance evaluation. Each model was optimized by identifying the best-performing hyperparameter configurations using a grid search strategy. Details of the explored hyperparameter search ranges and the best-performing configurations obtained through grid search are provided in Table 1.

Table 1. The explored hyperparameter search ranges and the best-performing configurations.

Model	**Hyperparameter Ranges**	**Best-performing Configuration**
Random Forest	max_depth: [None, 10, 20, 30]	max_depth: None
	n_estimators: [100, 200, 300]	n_estimators: 100
	min_samples_leaf: [1, 2, 4]	min_samples_leaf: 1
	min_samples_split: [2, 5, 10]	min_samples_split: 2
	class_weight: ['balanced_subsample', 'balanced']	class_weight: 'balanced'
XGBoost	colsample_bytree: [0.7, 0.8, 0.9]	colsample_bytree: 0.9
	learning_rate: [0.01, 0.1, 0.2]	learning_rate: 0.1
	gamma: [0, 0.1, 0.2]	gamma: 0.2
	max_depth: [3, 4, 5]	max_depth: 5
	subsample: [0.7, 0.8, 0.9]	subsample: 0.7
	n_estimators: [100, 200, 300]	n_estimators: 100
LightGBM	max_depth: [3, 5, 7]	max_depth: 3
	learning_rate: [0.05, 0.1]	learning_rate: 0.1
	n_estimators: [100, 200]	n_estimators: 200
	num_leaves: [15, 31, 63]	num_leaves: 15
ANN	optimizer: ['adam', 'rmsprop']	optimizer='adam’
	dropout_rate: [0.2, 0.3]	dropout_rate=0.3
	neurons: [32, 64]	neurons=32
	learning_rate: [0.001, 0.01]	learning_rate=0.001
	batch_size: [32, 64]	batch_size=32
	epochs: [100, 200]	epochs=200

Five-fold cross-validation was conducted on the training set to guarantee the models' robustness and reliability. During the cross-validation process, stratified splitting was performed to maintain the original class distribution across all folds, thereby addressing the class imbalance and ensuring that each fold contained a representative proportion of both good and bad credit cases. To

evaluate each model's capacity to generalize to new data, the mean and standard deviation of the evaluation metrics were computed across the five folds.

3.4. Evaluation and Interpretation

In this phase, several evaluation metrics are employed to assess the models' performance, including accuracy, precision, recall, F1-score, and AUC-ROC. These indicators offer a comprehensive view of how well the model predicts credit risk. To further support the robustness of the findings, paired t-tests are conducted on the F1 scores obtained from 5-fold cross-validation to determine whether the observed performance differences between models are statistically significant.

To enhance the interpretability and transparency of the model's predictions, XAI techniques, specifically SHAP and LIME, are utilized. By quantifying the contribution of each feature to the model output and offering instance-level explanations, SHAP provides a unified approach to both global and local interpretability. This enables stakeholders to understand which features are most influential across the entire dataset and how they impact individual predictions.

Meanwhile, LIME complements SHAP by generating local surrogate models to explain the behavior of the classifier for specific instances. It highlights the most influential features driving a particular prediction, allowing users to gain intuitive, human-understandable insights into the rationale for the model's decisions. The combined use of SHAP and LIME strengthens trust in the model by providing consistent and interpretable explanations at both the global and individual levels.

4. Experiment Results

4.1. Model Performance Comparison and Analysis

Table 2 presents the cross-validation performance metrics of four credit scoring models on the training set, including Random Forest, XGBoost, LightGBM, and ANN. Overall, while LightGBM excelled in AUC, the ANN model demonstrated more consistency and achieved the best average performance across all key metrics. This suggests that the ANN model consistently balanced identifying good credit and catching all true positives, resulting in the highest F1 score.

Table 2. Cross-validation metrics (mean ± std) of models for the training set. The best performances are bolded.

Model	Accuracy	Precision	Recall	F1 score	AUC
Random Forest	0.7637±0.0504	0.7203±0.0668	0.6813±0.0657	0.6920 ± 0.0680	0.7829 ± 0.0634
XGBoost	0.7450±0.0535	0.6927±0.0695	0.6714±0.0678	0.6783 ± 0.0693	0.7808 ± 0.0615
LightGBM	0.7625±0.0556	0.7206±0.0694	0.7006±0.0619	0.7066 ± 0.0654	**0.9544 ± 0.0912**
ANN	**0.7675±0.0591**	**0.7257±0.0668**	**0.7315±0.0774**	**0.7260 ± 0.0703**	0.7666 ± 0.0850

After training, the best-performing model from cross-validation was selected and evaluated on the holdout test set to assess its generalization ability and performance on unseen real-world data. Table 3 presents the performance metrics of four credit scoring models including Random

Forest, XGBoost, LightGBM, and Artificial Neural Network (ANN). The ROC curves for these models on the test dataset are illustrated in Fig. 2. Among them, the ANN model continuously performed better than the other models across all key evaluation metrics.

Table 3. Performance comparison of the models for the test set. The best performances are bolded.

Model	Accuracy	Precision	Recall	F1 score	AUC
Random Forest	**0.7250**	0.6671	0.6512	0.6571	0.7538
XGBoost	0.6700	0.6126	0.6167	0.6143	0.7175
LightGBM	0.7050	0.6538	0.6607	0.6567	0.7301
ANN	**0.7250**	**0.6734**	**0.6750**	**0.6742**	**0.7606**

The ANN achieved the highest accuracy of 0.7250, precision of 0.6734, recall of 0.6750, F1 score of 0.6742, and AUC of 0.7606. These results indicate that the ANN model not only correctly identified a greater number of true positive creditworthy individuals but also maintained a strong balance between precision and recall, reflecting robustness in both detection and error control.

By contrast, while the Random Forest model also achieved an accuracy of 0.7250, it fell slightly behind in other metrics, particularly in recall of 0.6512 and F1 score of 0.6571. XGBoost showed the weakest performance overall, with the lowest AUC of 0.7175 and F1 score of 0.6143, indicating a lower discriminative power. LightGBM performed moderately well, especially in recall of 0.6607.

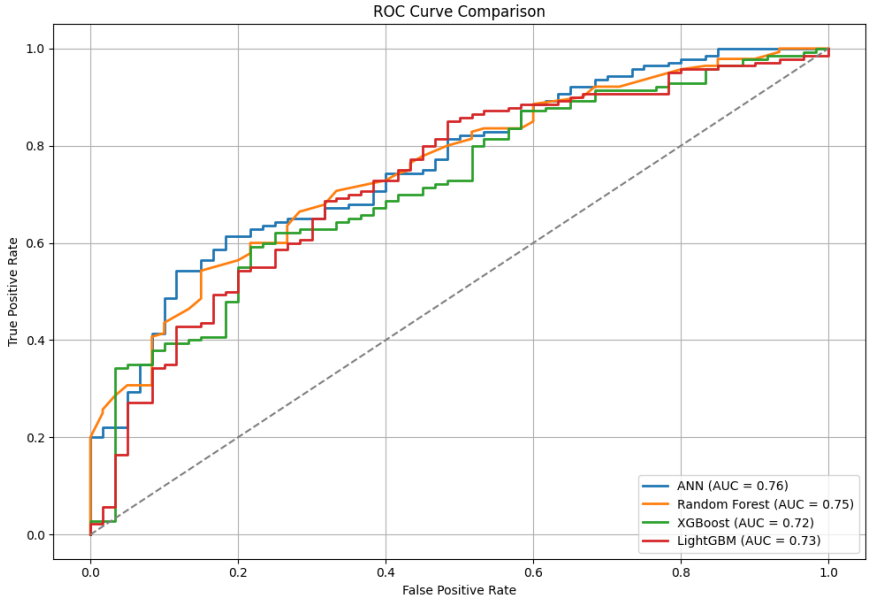

Fig. 2. Credit scoring models' ROC curves for the test set.

The ROC curves further reinforce this ranking, with the ANN model's curve lying above those of the other models, confirming superior classification capability across various threshold levels. Overall, the ANN demonstrated the most effective and reliable performance for credit scoring tasks in this experiment.

4.2. Significance Testing

To determine whether the performance improvements of the ANN model over other models are statistically significant, paired t-tests were conducted using F1 scores from 5-fold cross-validation. Table 4 provides a summary of the findings.

Table 4. The paired t-test results compare the ANN model to the other models.

Model Comparison	t-statistic	p-value
ANN vs Random Forest	3.62	0.0223
ANN vs XGBoost	2.76	0.0508
ANN vs LightGBM	4.58	0.0102

The comparisons between ANN and both Random Forest and LightGBM yielded p-values of 0.0223 and 0.0102, respectively, indicating statistically significant differences at the 5% significance level. These findings suggest that the ANN model significantly outperforms both Random Forest and LightGBM in terms of F1 score. When compared to XGBoost, the resulting p-value of 0.0508 indicates marginal significance, suggesting a potential performance advantage for the ANN model, although further data may be required to confirm this conclusively.

Overall, these statistical results provide additional rigor to the performance analysis and reinforce the claim that the ANN model is the most effective credit scoring approach among those evaluated in this study.

4.3. Interpretability Results Analysis

4.3.1. SHAP

To interpret the ANN model's predictions, SHAP was employed to evaluate each feature's global impact. The top ten variables on the German dataset that are important to the model prediction are shown in Fig. 3. It reveals that loan duration is the most influential factor affecting the model's output. The next most impactful features are checking_account_A14 (indicating the absence of a checking account) and checking_account_A11 (low balance in checking account), both of which suggest that clients with limited or no financial reserves are more likely to default. Additionally, credit_history_A34, which represents a critical or poor credit history, and savings_account_A61, indicating moderate savings, also substantially influence the model’s decision-making process. Other features, like age, credit purpose, installment rate, and telephone access, were found to have moderate influence.

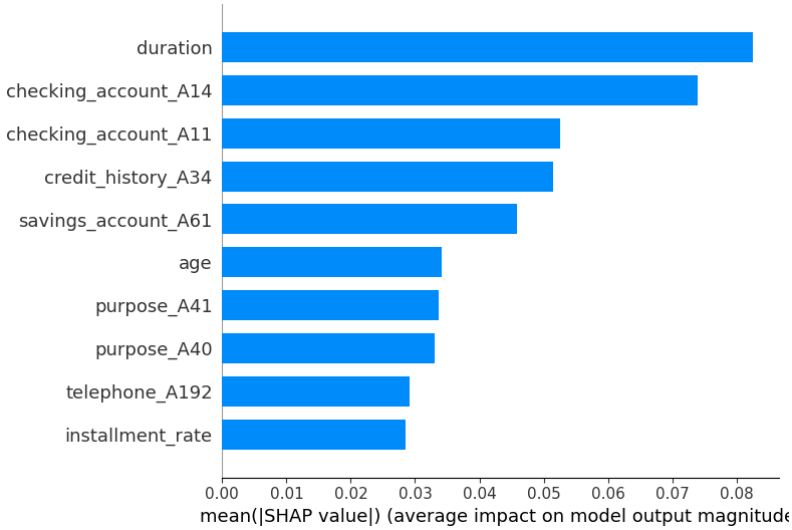

Fig. 3. SHAP global feature importance plot.

Fig. 4 presents the SHAP summary plot, which offers a local interpretability result for feature contributions and their respective values across all instances in the dataset. It reveals how feature values influence individual credit scoring outcomes. For example, longer loan durations generally increase SHAP values, pushing predictions toward bad credit classification indicating that longer credit durations are strongly associated with increased default risk. Similarly, clients with no or minimal checking account balances (e.g., checking_account_A14) tend to have positive SHAP values, meaning they significantly increase the model's confidence in predicting default. The plot also reveals features with bidirectional impact: such as, age and savings_account_A61 show both positive and negative SHAP values, indicating that their influence on credit risk is conditional.

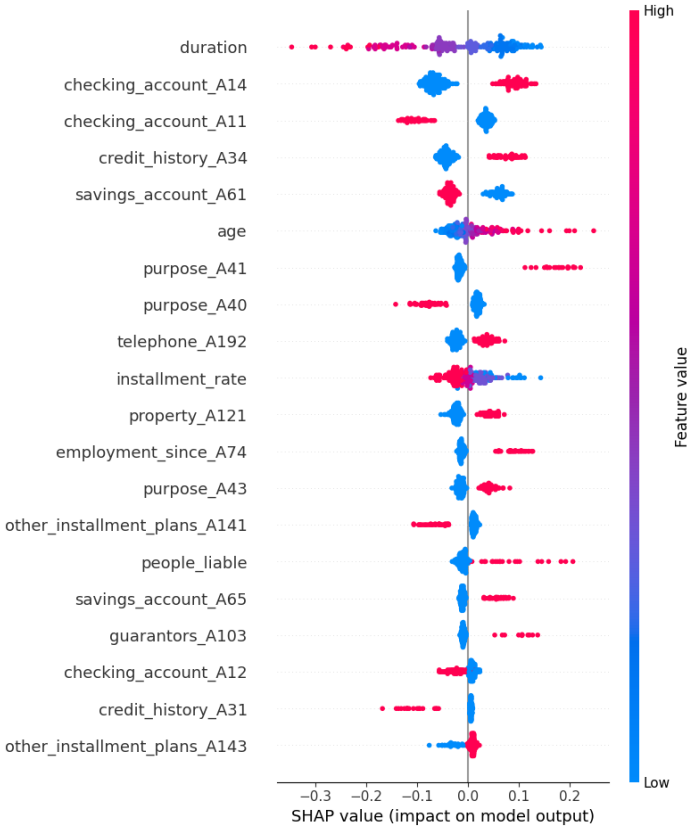

Fig. 4. SHAP summary plot.

Overall, this summary plot provides deeper insight into which client characteristics consistently raise red flags for default and which can potentially mitigate perceived risk, supporting more transparent and data-driven credit decisioning.

Fig. 5 presents the SHAP waterfall plot for an instance (i=20), illustrating how the model arrived at a final predicted probability of 0.801 for being a good credit. The base value, or average model output across the dataset, is approximately 0.645, and individual feature contributions either increase or decrease this base value to reach the final output. The most influential positive contributors pushing the prediction higher toward "good credit" include a relatively short loan duration (−0.727) contributing +0.07, the presence of a registered telephone line (telephone_A192 = 1) adding +0.06, and a low installment rate (−0.896) contributing +0.05. Additionally, age (0.758), purpose_A43 (e.g., used car), and checking_account_A11 (moderate balance) contribute positively to the credit score estimate.

In contrast, several features helped reduce the model's confidence in assigning a good credit label. These include checking_account_A14 = 0 and credit_history_A34 = 0 (no recent missed payments), which decreased the prediction by −0.05 and −0.04, respectively. Features like savings_account_A61, property_A12, and housing_A152 had smaller negative impacts, also slightly pulling the prediction downward.

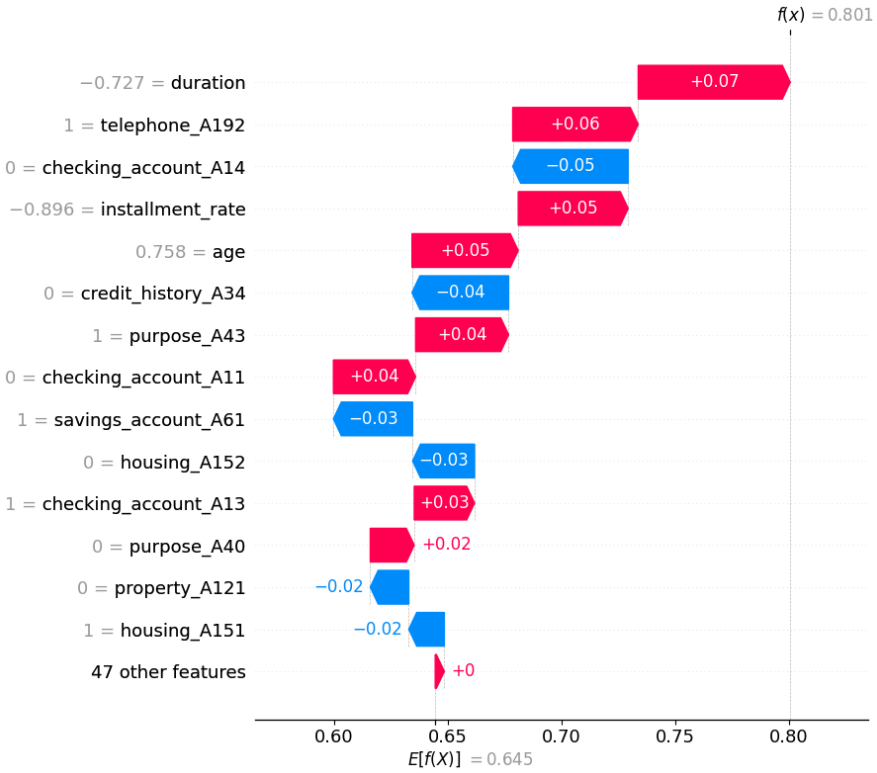

Fig. 5. SHAP waterfall plot for instance 20.

Overall, this plot gives a detailed explanation of how the model makes decisions for this specific client. It highlights that despite a few favorable factors, the presence of multiple risk indicators outweighed them, leading the model to strongly classify this instance as high risk.

4.3.2. LIME

To further interpret the model's decision, we employed LIME, which provides a linear surrogate model that approximates the classifier's behavior around the specific instance. As shown in Fig. 6, LIME predicted a high probability of 0.80 for being a good credit applicant for instance 20.

LIME highlights several key features that strongly influenced the outcome. Among the most supportive features for a good credit classification were the short loan duration (≤ −0.73), which contributed +0.21, the absence of a purpose_A41 loan (e.g., no new car loan), and the foreign_worker_A202 = 0 (likely indicating domestic worker status). These features are consistent with responsible financial behavior and thus increased the likelihood of being classified as low-risk.

On the other hand, features such as checking_account_A11 = 0, credit_history_A34 = 0, and guarantors_A103 = 0 slightly reduced the predicted probability by contributing towards the "poor credit" classification.

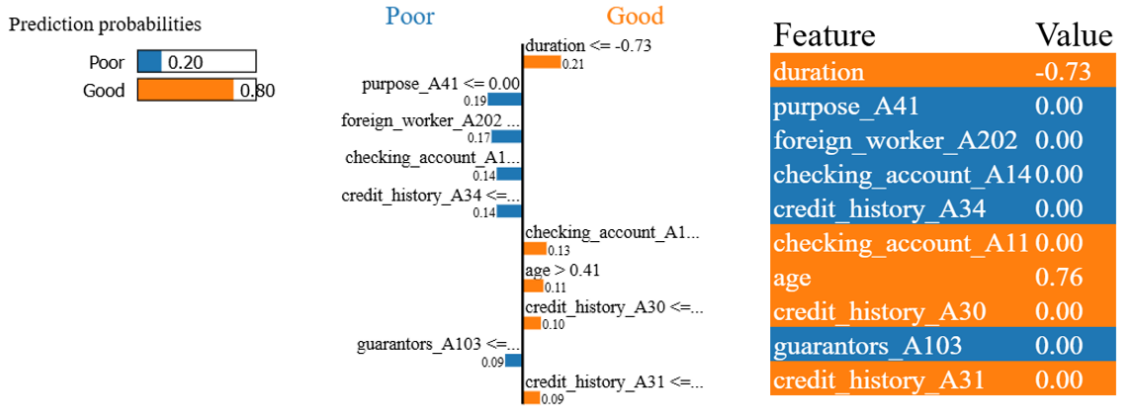

Fig. 6. LIME explanation for instance 20.

When compared to the SHAP explanation, LIME and SHAP show a strong level of agreement in identifying influential features. Both methods emphasize the importance of loan duration, checking account status, and credit history. Notably, both explanations identify a short loan duration and registered telephone line as significant positive indicators, while credit history (A34, A31) and low account balance (A14) act as moderating factors.

5. Discussion and Conclusion

This study investigated the application of deep learning models for credit scoring using the German Credit dataset, with a particular emphasis on model performance and interpretability through explainable AI (XAI) techniques. Among the models evaluated including Random Forest, XGBoost, LightGBM, and ANN, the ANN model consistently achieved superior results in both cross-validation and holdout test evaluations. Specifically, the ANN obtained the highest AUC of 0.7606, F1 score of 0.6742, recall of 0.6750 and precision of 0.6734 on the test set, demonstrating its strong capability in accurately identifying good credit applicants while minimizing false negatives. Results from the paired t-test indicated statistically significant differences between ANN and both Random Forest ($p = 0.0223$) and LightGBM ($p = 0.0102$), confirming ANN's superior performance at the 5% significance level.

In addition, SHAP and LIME were applied to interpret the ANN model's predictions. The results revealed that features such as loan duration, checking account status, credit history, age, and

savings account type were among the most influential factors affecting credit predictions. Notably, shorter loan durations were associated with higher predicted probabilities of good credit, while certain credit history and account types had mitigating effects. The consistency between SHAP and LIME not only strengthens the trustworthiness of the ANN model's predictions but also offers transparent, instance-specific explanations that can assist financial institutions in understanding and justifying automated credit decisions. The interpretability insights derived from SHAP and LIME allow stakeholders to audit model behavior, understand feature contributions, and mitigate potential biases.

In conclusion, the ANN model, coupled with explainable AI methods, offers a robust and interpretable solution that enhances both transparency and usability in financial decision-making. It not only delivers high predictive accuracy in classifying applicants as good or bad credit risks but also provides clear, human-understandable explanations for each decision. By making the reasoning behind each prediction clear, the model empowers financial organizations to make more responsible and informed loan decisions, while also helping borrowers understand the factors influencing their credit outcomes. Future research may consider incorporating more advanced neural architectures, enriching the dataset with additional socioeconomic variables, or applying counterfactual explanations to further improve decision support in financial services.

REFERENCES

Chen, Y., & Zhang, R. (2021). Research on credit card default prediction based on k-means SMOTE and BP neural network. *Complexity*, *2021*(1), 6618841.

Durand, D. (1941). Risk elements in consumer instalment financing. *Nber Books*.

GiciÄ‡, A., Ä©onko, D. e., & Subasi, A. (2023). Intelligent credit scoring using deep learning methods. *Concurrency and Computation: Practice and Experience*, *35*(9), e7637.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Luo, C., Wu, D., & Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, *65*, 465-470.

Museba, T. (2022). Adaptive Particle Swarm Optimized XGBoost Ensemble Algorithm for Online Credit Scoring.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,

Shen, F., Zhao, X., Kou, G., & Alsaadi, F. E. (2021). A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Applied Soft Computing*, *98*, 106852.

Talaat, F. M., Aljadani, A., Badawy, M., & Elhosseini, M. (2024). Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction. *Neural Computing and Applications*, *36*(9), 4847-4865.

Zhu, B., Yang, W., Wang, H., & Yuan, Y. (2018). A hybrid deep learning model for consumer credit scoring. 2018 international conference on artificial intelligence and big data (ICAIBD).

Zou, Y., Xia, M., & Lan, X. (2025). Interpretable credit scoring based on an additive extreme gradient boosting. *Chaos, Solitons & Fractals*, *194*, 116216.